Analyse Canonique Généralisée de données séquentielles

Romain Bar, Jean-Marie Monnez

ANALYSE CANONIQUE GÉNÉRALISÉE DE DONNÉES SÉQUENTIELLES

Romain Bar [1] & Jean-Marie Monnez [2]

[1] *Institut Elie Cartan, UMR 7502, Université de Lorraine, CNRS, INRIA*
BP 239, 54506 Vandoeuvre-lès-Nancy Cedex, France
Romain.Bar@iecn.u-nancy.fr
[2] *Jean-Marie.Monnez@iecn.u-nancy.fr*
http ://www.iecl.u-nancy.fr

Résumé. On suppose que des vecteurs de données arrivant séquentiellement dans le temps sont des observations i.i.d. d'un vecteur aléatoire. Après avoir défini un processus d'approximation stochastique de type Robbins-Monro de l'inverse d'une matrice de covariance, on définit une méthode récursive d'estimation séquentielle de vecteurs directeurs des r premiers axes principaux de l'analyse canonique généralisée de ce vecteur aléatoire. On étudie ensuite le cas où l'espérance de ces observations varie dans le temps. On donne finalement des résultats de simulation.

Mots-clés. Données de grande dimension, analyse de données séquentielles, analyse canonique généralisée, approximation stochastique. ...

Abstract. High dimensional data of a generalized canonical correlation analysis (gCCA) are supposed to be i.i.d. observations of a random vector Z which are taken sequentially. After defining a stochastic approximation process of the Robbins-Monro type to estimate sequentially the inverse of a covariance matrix, a recursive method of sequential estimation of unit vectors of the principal axes of gCCA is defined. Next, the case where the expectation of the n^{th} observation varies with time n is studied. Finally, simulation results are given.

Keywords. High dimesional data, sequential data analysis, generalized canonical correlation analysis, stochastic approximation. ...

1 Introduction

On observe p caractères quantitatifs sur des individus : on obtient des vecteurs de données z_i dans $\mathbb{R}^p$. On se place ici dans le cas où ces vecteurs arrivent séquentiellement dans le temps : on observe z_n au temps n ; on a une suite de vecteurs de données $z_1, \ldots, z_n, \ldots$.

On suppose que $z_1, \ldots, z_n, \ldots$ constituent un échantillon i.i.d. d'un vecteur aléatoire Z dans $\mathbb{R}^p$ défini sur un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$. Ω représente une population d'où on extrait un échantillon.

On se place dans le cas où le vecteur aléatoire Z est partitionné en sous−vecteurs $Z^1,\ldots,Z^q$; pour $k=1,\ldots,q$, Z^k est un vecteur aléatoire dans $\mathbb{R}^{m_k}$, de composantes $Z^{k1},\ldots,Z^{km_k}$. On souhaite effectuer une ACP de Z dans laquelle les vecteurs aléatoires Z^k aient un rôle équilibré : on veut éviter que les premiers facteurs soient principalement déterminés à partir de certains vecteurs Z^k. L'analyse canonique généralisée du vecteur aléatoire Z (ACGVA)fournit une solution à ce problème.

L'ACGVA représente l'ACG effectuée sur la population Ω dont on va chercher à estimer au temps n les résultats à partir des données dontt on dispose à ce temps. Soit θ un résultat de l'ACGVA, par exemple une valeur propre, un facteur (c'est ce dernier cas qui est considéré ici).

Plutôt que d'effectuer à chaque temps n une estimation de θ à partir de l'ensemble des données dont on dispose jusqu'à ce temps, on va effectuer une estimation récursive de θ : disposant d'une estimation θ_n de θ obtenue à partir des observations $z_1,\ldots,z_{n-1}$, on introduit au temps n l'observation z_n et on définit à partir de θ_n et z_n une nouvelle estimation θ_{n+1} de θ :

$$\theta_{n+1} = f_n(\theta_n, Z_n)$$

.

On utilise pour cela un processus d'approximation stochastique.

2 ACG d'un vecteur aléatoire

On suppose qu'il n'existe pas de relation affine entre les composantes du vecteur aléatoire Z. Le critère de l'ACG est le suivant : pour $l=1,\ldots,r$, déterminer au pas l une combinaison linéaire des composantes centrées de Z, $U_l=\theta_l'(Z-\mathbb{E}[Z])$, et pour $k=1,\ldots,q$, une combinaison linéaire des composantes centrées de Z^k, $V_l^k=(\eta_l^k)'(Z^k-\mathbb{E}[Z^k])$, telles que :

$$\begin{array}{rcl} \sum_{k=1}^{q}\rho^2 \quad (U_l,V_l^k) & \quad \max & \\ Var(U_l) & = & 1, \\ Cov(U_l,U_j) & = & 0,\ j=1,\ldots,l-1, \\ Var(V_l^k) & = & 1;,k=1,\ldots,q. \end{array}$$

Soit C la matrice de covariance de Z, C^k celle de Z^k et M la métrique diagonale par blocs dans $\mathbb{R}^p$:

$$M = \begin{pmatrix} (C^1)^{-1} & & & \\ & \cdot & & \\ & & \cdot & \\ & & & (C^q)^{-1} \end{pmatrix}$$

θ_l, appelé l $_{\text{ième}}$ facteur général, est vecteur propre de la matrice MC associé à la $l_{\text{ième}}$ plus grande valeur propre. On peut interpréter ce résultat de la façon suivante : θ_l est le $l_{\text{ième}}$ facteur de l'ACP de Z dans $\mathbb{R}^p$ muni de la métrique M. $v_l = M^{-1}\theta_l$ est un vecteur directeur du $l_{\text{ième}}$ axe principal de cette ACP, vecteur propre de CM. Dans le cas particulier où, pour tout k, Z^k est de dimension 1, on retrouve l'ACP normée.

3 Approximation stochastique des vecteurs v_l

Pour définir le processus d'approximation stochastique, on a besoin au pas n d'estimateurs convergents M_n de M, obtenus à partir des observations $Z_1, \ldots, Z_{n-1}$.

Soit le vecteur aléatoire Z_1^k de dimension $m_k + 1$, obtenu en ajoutant au vecteur Z^k une dernière composante égale à 1. Soit J la matrice $(m_k + 1, m_k)$ obtenue en ajoutant à la matrice-identité d'ordre m_k une dernière ligne de zéros. on établit que la matrice $(m_k + 1, m_k)$,

$$X^k = \begin{pmatrix} (C^k)^{-1} \\ -(\mathbb{E}[Z^k])'(C^k)^{-1} \end{pmatrix}$$

,

est solution de l'équation en X : $\mathbb{E}[Z_1^k(Z_1^k)'X - J] = 0$.

On définit récursivement le processus (M_{1n}^k) d'approximation stochastique de X^k de type Robbins-Monro dans l'ensemble des matrices $(m_k + 1, m_k)$:

$$\begin{aligned} M_{1,n+1}^k &= M_{1n}^k - a_n(Z_{1n}^k(Z_{1n}^k)'M_{1n}^k - J), \\ a_n \geq 0, &\quad \sum_1^\infty a_n = \infty, \sum_1^\infty (a_n)^2 \leq \infty. \end{aligned}$$

Soit M_n^k la matrice obtenue à partir de M_{1n}^k en enlevant la dernière ligne ; on définit comme estimateur de M au pas n la matrice diagonale par blocs M_n qui a pour $k_{\text{ième}}$ bloc diagonal M_n^k.

Un estimateur de $\mathbb{E}[Z]$ au pas n est la moyenne empirique $\overline{Z}_{n-1}$ des observations $Z_1,\ldots,Z_{n-1}$, que l'on calcule récursivement.

Le vecteur directeur v_l du $l_{\text{ième}}$ axe principal de l'ACG est vecteur propre de la matrice M-symétrique $B = CM = (\mathbb{E}[ZZ'] - \mathbb{E}[Z]\mathbb{E}[Z'])M$ associé à la $l_{\text{ième}}$ plus grande valeur propre.

La fonction $F(x) = \frac{<Bx,x>_M}{||x||_M^2}$ est maximale pour $x = v_1$ et vaut alors λ_1 ; sous la contrainte que x est M-orthogonal à $v_1,\ldots,v_{l-1}$, elle est maximale pour $x = v_l$.

En suivant Bouamaine et Monnez (1998), on définit récursivement un processus d'approximation stochastiquue $(X_n) = ((X_n^1,\ldots,X_n^r))$ de $(v_1,\ldots,v_r)$:

$$\begin{aligned} B_n &= (Z_nZ_n' - \overline{Z}_{n-1}\overline{Z}_{n-1}')M_n, \\ F_n(X_n^l) &= \frac{<B_nX_n^l,X_n^l>_{M_n}}{||X_n^l||_{M_n}^2}, \\ Y_{n+1}^l &= X_n^l + \frac{a}{n^\alpha}(B_n - F_n(X_n^l)I)X_n^l,\ l = 1,\ldots,r \\ X_{n+1} &= \text{orth}_{M_n}(Y_{n+1}). \end{aligned}$$

Pour obtenir X_{n+1}, on effectue une orthogonalisation au sens de Gram-Schmidt par rapport à M_n de $Y_{n+1} = (Y_{n+1}^1,\ldots,Y_{n+1}^r)$.

On établit la convergence de ce processus pour $\frac{2}{3} \leq \alpha \leq 1$.

On peut définir des variantes de ce processus en utilisant au pas n, au lieu d'une seule observation Z_n, plusieurs observations de Z, ou toutes les observations faites jusqu'à ce pas, en remplaçant dans la définition de B_n le produit Z_nZ_n' par la moyenne des produits Z_iZ_i' utilisés à ce pas.

4 Cas où l'espérance des observations varie dans le temps

On suppose que, pour tout $n \geq 1$, l'espérance mathématique de Z_n, θ_n, dépend du temps n : $Z_n = \theta_n + R_n$, les R_n constituant un échantillon i.i.d d'un vecteur aléatoire R d'espérance nulle et de matrice de covariance C.
On note $\theta_n^k = \mathbb{E}[Z_n^k],\ k = 1,\ldots,q$.

L'ACG de R est appelée ACG partielle.
Les vecteurs directeurs v_l des axes principaux sont vecteurs propres de CM, avec $C = \mathbb{E}[(Z_n - \theta_n)Z_n']$ pour tout $n \geq 1$, M étant la métrique définie dans le paragraphe 2, avec $C^k = \mathbb{E}[(Z_n^k - \theta_n^k)(Z_n^k)']$, pour tout $n \geq 1$, pour $k = 1,\ldots,q$.

On suppose que l'on dispose au temps n d'un estimateur Θ_n de θ_n (ou, pour $k = 1, \ldots, q$ d'un estimateur Θ_n^k de θ_n^k) vérifiant certaines hypothèses.

Pour $k = 1, \ldots, q$, $(C^k)^{-1}$) est solution de l'équation en X : $\mathbb{E}[(Z_n^k - \theta_n^k)(Z_n^k)'X - I] = 0$ où I est la matrice-identité d'ordre m_k.

On définit récursivement le processus d'approximation stochastique de $(C^k)^{-1}$, (M_n^k), par :

$$M_{n+1}^k = M_n^k - a_n((Z_n^k - \Theta_n^k)(Z_n^k)'M_n^k - I)$$

.

On définit comme estimateur de M au pas n la matrice diagonale par blocs M_n qui a pour $\mathrm{k}_{\text{ième}}$ bloc diagonal M_n^k.

On définit récursivement un processus d'approximation stochastique $(X_n) = ((X_n^1, \ldots, X_n^r))$ de $(v_1, \ldots, v_r)$ par :

$$\begin{aligned} B_n &= (Z_n Z_n' - \Theta_n \Theta_n')M_n, \\ F_n(X_n^l) &= \frac{<B_n X_n^l, X_n^l>_{M_n}}{||X_n^l||_{M_n}^2}, \\ Y_{n+1}^l &= X_n^l + \frac{a}{n^\alpha}(B_n - F_n(X_n^l)I)X_n^l,\ l = 1, \ldots, r \\ X_{n+1} &= \mathrm{orth}_{M_n}(Y_{n+1}). \end{aligned}$$

Un cas particulier de ce modèle d'évolution de l'espérance θ_n de Z_n dans le temps est le suivant. Si l'on note θ_n^i la $i_{\text{ème}}$ composante réelle de θ_n $(i = 1, \ldots, p)$, on définit le modèle linéaire $\theta_n^i =< \beta^i, U_n^i >$, $< \cdot, \cdot >$ désignant le produiit scalaire euclidien usuel dans $\mathbb{R}^{n_i}$, U_n^i étant un vecteur de valeurs au temps n de fonctions connues du temps ou un vecteur de valeurs de variables explicatives contrôlées et β^i un vecteur inconnu de $\mathbb{R}^{n_i}$.

On définit alors comme dans Monnez (2008b), pour $i = 1, \ldots, p$, le processus d'approximation stochastique (B_n^i) de β^i et l'estimation Θ_n^i de θ_n^i par :

$$\begin{aligned} B_{n+1}^i &= B_n^i - a_n U_n^i((U_n^i)'B_n^i - Z_n^i), \\ \Theta_n^i &= < B_n^i, U_n^i > \end{aligned}$$

Z_n^i étant la $i_{\text{ème}}$ composante réelle de Z_n.

5 Mise en oeuvre et conclusion :

Les simulations ont été effectuées avec le logiciel R et consistent à tester la rapidité de notre méthode pour calculer les facteurs de l'ACG dans le cas où les observations sont i.i.d. L'idée

générale du programme est la suivante :

1) Initialisation : on prend en compte un petit nombre d'observations afin de calculer une première valeur pour la matrice de covariance empirique, C_0, la métrique, M_0, et les facteurs.

2) Pas n : On introduit une nouvelle donnée,

a) d'une part, on met à jour la matrice de covariance empirique, la métrique à l'aide d'une formule récursive : on obtient alors les estimations C_n et M_n. On calcule alors grâce à la routine Lapack utilisée dans R (, Scilab, ...) les r premiers vecteurs propres de la matrice M_nC_n qui sont des estimations des facteurs de l'ACG correspondants.

b) d'autre part, on met en oeuvre les différents processus décrits dans le paragraphe 2 pour finalement obtenir, pour $l = 1, \ldots, r$, X_n^l, estimation d'un vecteur directeur du $l_{\text{ième}}$ axe principal. On obtient une estimation des r premiers facteurs de l'ACG grâce à la relation : $\theta_n^l = M_n X_n^l$.

On choisit le temps durant lequel tourne l'algorithme (en supposant que le flux de données est continu) ainsi que la taille du vecteur Z, on compare alors la précision des 2 méthodes via la valeur du cosinus de l'angle formé par les facteurs théoriques et calculés.

La méthode développée dans l'article s'avère particulièrement efficace dans le cas où le vecteur Z est de grande dimension.

Un prolongement est d'implémenter la méthode développée dans le paragraphe 2 où l'espérance des observations varie dans le temps. On peut aussi étudier le cas où la matrice de covariance des observations varie aussi dans le temps.

Bibliographie

[1] Benzecri, J.P. (1969), Approximation stochastique dans une algèbre normée non commutative, Bulletin de la SMF, 97 :225-241.
[2] Bouamaine, A. et Monnez, J.M. (1998), Approximation stochastique de vecteurs et valeurs propres, Publications de l'iSUP, 42, n°2-3 :15-38.
[3] Monnez, J.M. (2008a), Stochastic approximation of the factors of a generalized canonical correlation analysis, Statistics & Probability Letters, 78 :2210-2216.
[4] Monnez, J.M. (2008b), Analyse en composantes principales d'un flux de données d'espérance variable dans le temps, RNTI, C-2 :43-56.
[5] Robbins, H. et Monro, S. (1951), A stochastic approximation method, AMS, 22 :400-407